Computational Tools for the Identification and Interpretation of Sequence Motifs in Immunopeptidomes

Alvarez, Bruno; Barra, Carolina; Nielsen, Morten; Andreatta, Massimo

Published in:
Proteomics

Link to article, DOI:
10.1002/pmic.201700252

Publication date:
2018

Document Version
Peer reviewed version

Computational Tools for the Identification and Interpretation of Sequence Motifs in Immunopeptidomes

Article *in* Proteomics · January 2018

DOI: 10.1002/pmic.201700252

CITATIONS
13

READS
136

4 authors, including:

Bruno Alvarez
National University of General San Martín
16 PUBLICATIONS **53** CITATIONS
SEE PROFILE

Morten Nielsen
Technical University of Denmark
369 PUBLICATIONS **11,786** CITATIONS
SEE PROFILE

Massimo Andreatta
National University of General San Martín
49 PUBLICATIONS **1,039** CITATIONS
SEE PROFILE

Computational Tools for the Identification and Interpretation of Sequence Motifs in Immunopeptidomes

Bruno Alvarez[1*], Carolina Barra[1*], Morten Nielsen[1,2], Massimo Andreatta[1]

[1] Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, CP1650 San Martín, Argentina

[2] Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark

* Both authors contributed equally to this work

Corresponding author: Massimo Andreatta, PhD [mandreatta@iibintech.com.ar] Instituto de Investigaciones Biotecnológicas, Universidad Nacional de San Martín, Campus Miguelete, Av. 25 de Mayo y Francia, CP1650 San Martín, Argentina

Keywords: GibbsCluster, mass spectrometry, MHC, prediction models, sequence motifs

Word count: 7443

Abstract

Recent advances in proteomics and mass-spectrometry have widely expanded the detectable peptide repertoire presented by major histocompatibility complex (MHC) molecules on the cell surface, collectively known as the immunopeptidome. Finely characterizing the immunopeptidome brings about important basic insights into the mechanisms of antigen presentation, but can also reveal promising targets for vaccine development and cancer immunotherapy. In this report, we describe a number of practical

Received: 10 06, 2017; Revised: 12 15, 2017; Accepted: 01 03, 2018

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/pmic.201700252.

and efficient approaches to analyze immunopeptidomics data, discussing the identification of meaningful sequence motifs in various scenarios and considering current limitations. We address the issue of filtering false hits and contaminants, and the problem of motif deconvolution in cell lines expressing multiple MHC alleles, both for the MHC class I and class II systems. Finally, we demonstrate how machine learning can be readily employed by non-expert users to generate accurate prediction models directly from mass-spectrometry eluted ligand data sets.

Introduction

The comprehensive set of peptides presented on the cell surface by MHC molecules, collectively referred to as the immunopeptidome, represents a unique fingerprint of the health of a cell. T lymphocytes routinely scan this pool of MHC-associated peptides, and can help eliminating infected or cancerous cells that present abnormal peptides on their surface. MHC class I molecules mainly bind peptides derived from intracellular pathogens (such as viruses and some bacteria) and present them to cytotoxic T lymphocytes; MHC class II epitopes are mainly derived from extracellular proteins and are presented to T-helper lymphocytes.

Recent technological advances in the field of mass spectrometry (MS) have brought about a revolution in the study of immunopeptidomes (reviewed in ref. [1]), with several thousands of peptides that can be detected in a single experiment. Large data sets of naturally presented peptides have been beneficial to define more accurately the rules of peptide-MHC binding [2–4] but have also a tremendous potential in defining pathogen-derived T cell epitopes [5,6] and neo-epitopes unique to cancerous cells [7–10]. Part of the appeal of MS-based approaches is that they do not require prior knowledge of MHC motifs, and there is no human intervention in defining a library of candidate sequences to be tested. Therefore, MS provides a large but relatively unbiased sampling of the population of processed and presented peptides available for T cell recognition [3].

In most MS-based pipelines, spectra from eluted peptides are matched against a reference database of natural proteins using algorithms like MaxQuant [11] or PEAKS [12,13], and filtered against a decoy database to limit the false discovery rate (FDR). Strict FDR filters (typically in the order of 1%) should ensure that most spectra are correctly assigned to *bona fide* ligands, but often leads to discarding a large portion of the spectra. Several approaches

have been proposed to increase the yield of spectral assignment. For example, Mascot Percolator performs machine learning on high-confidence matches to re-score database search results for lower-confidence peptides [14]. Instead of matching spectra to an entire protein database, SpectMHC constructs reduced, targeted databases of potential MHC ligands, effectively reducing the amount of spurious decoy hits [15]. Recent work has also suggested that a portion of the unassigned spectra may also be explained by proteasome-generated spliced peptides, which would require the inclusion of spliced variants in the target database [16,17].

After spectral assignment to amino acid sequences, peptides must often be aligned and/or clustered to extract meaningful sequence motifs of antigen presentation. The analysis protocols here will generally differ depending on the type of receptor (MHC I vs. MHC class II) and type of sample used (cellular versus soluble MHC molecules and mono- vs. poly-allelic cell lines). MHC I ligands have a limited range of lengths, typically 8 to 11 amino acids long, and are characterized by very conserved amino acid preferences at the positions interacting with the MHC binding groove (anchor positions). On the other hand, MHC II ligands are normally longer, with only a portion, the binding core, directly interacting with the MHC groove [18]; in this case a more sophisticated alignment process is needed to extract conserved binding preferences. In transgenic cells expressing a single MHC molecule (mono-allelic), only one specificity is expected to be present in the data and motif identification is relatively straightforward. Conversely, unmodified cells will naturally present peptides bound to multiple MHC alleles (up to six for HLA class I), with generally different binding preferences; in this case, the multiple specificities contained in the data must be deconvoluted, either by assigning MHC restriction with predictive methods, or by unsupervised clustering.

A popular tool for the unsupervised identification of sequence motifs in immunopeptidomes is GibbsCluster [19,20], a web-based and downloadable method that has been included into numerous pipelines for the deconvolution of ligand motifs in the MHC class I [21–23,10] and MHC class II [24–26] systems. The GibbsCluster algorithm takes as input a list of peptide sequences (potentially of variable length), and uses a heuristic search to group them into information-rich groups. Besides the sequence motif defining each group, additional properties such as the ligand length distribution of each cluster can be analyzed. A similar method, MixMHCp [2,27], has shown performance comparable to GibbsCluster, with the limitation that it can only handle peptides of uniform length. A useful feature of GibbsCluster is the “trash cluster”, a check on internal motif consistency that can filter out outliers that

cannot be assigned to any clusters. In the context of MS eluted ligand data, spurious data points can originate both from LC-MS/MS contaminants and from erroneous spectral matches. As a noise filter, GibbsCluster can be beneficial also for mono-allelic data sets where no motif deconvolution is required.

While sequence motifs are generated by GibbsCluster in an unsupervised manner, the method cannot directly assign the MHC restriction of each ligand; this must be done by comparing the unsupervised motifs with published binding motifs of the MHC molecules in the sample [28]. While this comparative approach is in most cases feasible for human MHC, whose most prevalent alleles have been well characterized and documented, it will fall short for samples containing uncharacterized specificities. Aiming to overcome this limitation, Bassani-Sternberg et al. [27] suggested a strategy for automatic, unbiased annotation of MHC restriction by comparing motifs detected in multiple data sets with known haplotypes. Exploiting the co-occurrence of MHC alleles across different data sets, they were able to assign motifs to individual alleles without relying on *a priori* assumptions on their binding specificity, also for alleles without previously documented ligands.

Over the past decades, many efforts have been dedicated to the development of computational methods for the prediction of peptide binding to MHC class I molecules. Most of these T-cell epitope prediction methods have been traditionally trained solely on *in-vitro* data of peptide-MHC binding affinity. Although peptide-MHC affinity is arguably the most selective step in antigen presentation, other factors influence the likelihood of a peptide being presented on the cell surface for T-cell recognition [29,30]. *In-vitro* binding affinity data does not address the fact that antigen presentation is a complex, integrative physiological process that combines antigen processing, transport and binding affinity/stability of the peptide-MHC complex. Finally, *in-vitro* data fails to reflect any peptide length preference of different MHC-I alleles. Because naturally eluted ligands incorporate information about these additional properties of antigen presentation, large MS-derived sets of peptides can potentially enable the generation of more accurate prediction models. Recent studies have suggested that models trained on MHC class I ligand data outperform binding affinity-based predictors when it comes to identification of eluted ligands and T cell epitopes, both in an allele-specific setting [2,3] as well as with pan-allelic coverage [4]. Generic tools for machine learning from peptide sequences such as NNAlign [31,32] can be applied to individual MS data sets to generate custom-made prediction models, which can in turn be employed for further downstream analyses of the immunopeptidome.

The rapidly expanding collection of naturally eluted ligands revealed by MS and the analysis toolkits developed in its wake hold great promise in understanding the structure of the immunopeptidome and the rules of antigen presentation. However, because of the complexities inherent to MS eluted ligand data, it is not a trivial task to analyze and interpret the information these data sets contain. In this report, we seek to address some common issues and describe strategies to analyze MS ligand data and derive sequence motifs in the various scenarios outlined above (MHCI vs. MHCII; mono-allelic vs. poly-allelic cell lines), with guidelines and examples on publicly available datasets.

MHC class I; mono-allelic cells

In a recent publication, Abelin et al. [3] described the development of transgenic cells that express a single human MHC class I allele (HLA), and used them to generate a large set of MHC ligands covering 16 HLA class I alleles. There are obvious advantages in using mono-allelic cells to characterize MHC ligands: firstly, no deconvolution/clustering is required to define motifs at the single-allele resolution; secondly, the assignment of individual peptides to their allele does not have to depend on binding predictions or prior knowledge of the motifs. Apart from technical difficulties in the cell generation, a possible drawback is that the relative level of expression of different MHC alleles in a given cell, and the amount of ligands they present, is lost in a mono-allelic setting. The amount of ligands presented by different alleles may also depend on competition between MHC molecules, where the newly available digested peptides from an unfolding antigen fragment would presumably be captured by MHCs with the highest affinity [33].

Although most software for MS spectra mapping uses a strict false discovery rate (FDR) threshold, incorrect ligands may still be present among the matches that pass the FDR check. These may consist of common contaminants such as keratin or histone proteins, as well as residual peptides from previous runs of the LC-MS/MS instruments used for sample preparation [34,35]. GibbsCluster is a useful tool to detect and remove such contaminants and false hits. For each allele in the Abelin data set [3], we applied GibbsCluster-2.0 with default preset options for “MHC class I ligands of length 8-13”, specifying a single cluster. Between 0.4% and 16% of the peptides (mean 4%) of length 8 to 13 were inconsistent with the motif identified by GibbsCluster-2.0 and were removed by the program as noise. While distinct motifs can be discerned before trash cluster filtering (see three representative alleles in Figure 1A), the post-filtering motifs have higher information content and more well-defined

anchor residues (Figure 1B). Peptides in the "trash cluster" may sometimes hint at the origin of the contamination: for example, the observation of terminal Arginine/Lysine preferences at the C-terminus in several of the 16 alleles points towards tryptic peptides polluting the mixtures (Supplementary Figure S1). The ligands in the Abelin data set have in general very good correspondence to known MHC binding preferences, with an average NetMHCpan-3.0 percentile rank [36] well below 1% for most alleles (Figure 1C, red boxplots). In contrast, peptides in the trash cluster match very poorly the preferences of their MHC and are assigned high NetMHCpan rank scores (Figure 1C, blue boxplots).

MHC class I; poly-allelic cells

Unmodified antigen-presenting cells will generally express up to six different MHC class I alleles (two each for HLA-A, HLA-B and HLA-C). The immunopeptidome of these cells therefore consists of multiple specificities mixed together, where the global haplotype is known but the restriction of each individual ligand is unknown. For example, Bassani-Sternberg et al. [21] described the LC-MS/MS analysis of peptides eluted from seven different cancer cell lines and primary cells, which had been HLA-typed at high resolution, and demonstrated how the GibbsCluster approach could be used to deconvolute the individual peptide restrictions. Here we illustrate the application of GibbsCluster to one of the cell lines from the Bassani-Sternberg study, HCC1143, which expresses the five alleles HLA-A*31:01, HLA-B*35:08, HLA-B*37:01, HLA-C*04:01, HLA-C*06:02.

GibbsCluster finds an optimal solution of four clusters, with a close correspondence to all but one of the HC1143 alleles (Figure 2), failing to separate HLA-C*04:01 ligands. HLA-C molecules have low expression levels and rather degenerate binding preferences [27,37], making the deconvolution of their motifs more challenging. The motifs determined by unsupervised clustering show a remarkable correspondence with the binding preferences predicted by NetMHCpan-3.0 [36]. There are, in some instances, subtle differences between the NetMHCpan and GibbsCluster motifs, as in the case of additional secondary anchors (e.g. a positively charged P5 for HLA-B*37:01). This suggests that motifs directly derived from eluted ligands may carry an additional level of information on peptide presentation (for instance, secondary anchors conferring improved peptide-MHC complex stability) compared to the NetMHCpan motifs, which were constructed from *in vitro* binding affinity data. The sizes of the clusters give an indication of the relative level of expression of the different alleles, with the largest group corresponding to the homozygous HLA-A*31:01 (1253

peptides), followed by the two HLA-B alleles (610 and 460 peptides respectively) and by the lowly-expressed HLA-C*06:02 (409 peptides). Finally, 45 peptides were collected by the trash cluster. Interestingly, for six out of seven cell lines in the Bassani-Sternberg data set, we noted a C-terminal enrichment for Arginine/Lysine in peptide discarded in the trash cluster (Supplementary Figure S2). A similar observation was made for the Abelin data set discussed previously, and hints that residual peptides derived from trypsin digestion may often be present in the LC column.

As an alternative approach to unsupervised clustering, one can assign each peptide to a MHC allele using peptide-MHC binding prediction methods; then deriving sequence motifs from the resulting groups of peptides. We applied NetMHCpan [36] to the peptides eluted from the HCC1143 cell line, assigning peptides to the MHC molecule in the haplotype with the lowest predicted NetMHCpan percentile rank. If a peptide could not be assigned to any MHC molecule with rank<=2%, then it was discarded in a trash cluster. While this setup mimics the GibbsCluster strategy, it has the very important difference that NetMHCpan utilizes known motif preferences of the MHC molecules to make the assignments, whereas GibbsCluster is unsupervised and requires no prior knowledge of the motifs. In the case of the HCC1143 cell line, the MHC molecules are all well characterized and the solutions found by the two approaches are remarkably similar (Supplementary Figure S3). Assignment by NetMHCpan has the potential advantage that at least a fraction of peptides could be assigned to HLA-C*04:01, a specificity that was not detected by unsupervised clustering. However, in cases where the haplotype is not fully characterized, or when the known MHC alleles have poorly studied motifs, the assignment by NetMHCpan would fail. This is exemplified by a recent study of bovine MHC ligands [38], for which the motifs derived by GibbsCluster differed dramatically from the assignments made by NetMHCpan due to paucity of training data available to NetMHCpan for these alleles. Note also, that the number of ligands discarded to the trash cluster using this approach was more than 10 times higher compared to those discarded by GibbsCluster (463 versus 45).

MHC class II, mono-allelic cells

Analyzing MHC class II binding data is for many reasons more complex compared to MHC class I. First and foremost, the HLA class II binding groove is open at both ends, accommodating peptides of a wide range of length by letting them protrude at either terminus of the nonamer binding core. Sophisticated alignment methods are therefore

required to identify the conserved binding preferences of MHC class II molecules [39–41]. Secondly, the binding motifs for MHC class II are in general more degenerate compared to the highly conserved MHC class I motifs [42,43]. These observations make the analysis and interpretation of MHC class II binding data, including MS ligands, highly challenging.

In a recent paper by Ooi et al. [44], MS eluted ligand data were used to investigate how patients expressing different HLA class II alleles have different susceptibility to autoimmune diseases. To characterize the specificity for each allele, they generated transgenic mice bearing the human HLA-DR1 MHC class II allele. On these data, we illustrate how the GibbsCluster method can be used to identify the binding motif of MHC class II molecules from mono-allelic MS ligand data and at the same time remove potential outliers. The 5740 non-redundant raw eluted peptide sequences were uploaded to the GibbsCluster web server, setting the recommended preset parameters for MHC class II peptides, except for the number of iterations per sequence per temperature step (set to 100) and the number of temperature steps (set to 50); these parameters entail a slower, but more accurate, motif search. The method recovered the binding motif for allele HLA-DRB1*01:01, with strong amino acid preferences at anchor residues at P1, P4, P6 and P9 (Figure 3A). These preferences were observed both without (Figure 3A, left panel) or with a trash cluster activated (Figure 3A, right panel). By activating the trash cluster option with a threshold of 2, 179 peptides (3% of data) were removed, and the logo showed a 20% increase in information content (Figure 3A, right panel).

MHC class II, poly-allelic cells

Another data set obtained from the Ooi et al. study [44] consists of peptides eluted from HW09013 cells that express the HLA-DR15/DR51 class II alleles. On this poly-allelic data set of MS eluted ligands, we set out to demonstrate how the GibbsCluster can be used to separate multiple specificities in MHC class II ligand data. The set of 2782 unique eluted peptides was submitted to GibbsCluster, using the recommended preset parameters for MHC class II and allowing the program to search up to three clusters. The unfiltered, single-cluster solution shows a motif with the correct P1, P4, P6 and P9 anchor positions, but with low information content and preferences that are a mixture of the two alleles in the sample (Figure 3B, left panel). Activating the trash cluster with a threshold of 2, the maximum information content is observed for the solution with two clusters (Figure 3B, right panel). The amino acid preferences identified by GibbsCluster resemble previously published motifs

derived from binding affinity data for HLA-DRB1*15:01 and HLA-DRB5*01:01 [31,45], and closely overlap with the global peptidome of DR15/51 characterized in a recent study [46]. Specifically, cluster 1 was composed of 1610 peptides (57.9%) and its motif resembles the HLA-DR15 binding preferences; cluster 2 comprised 1050 peptides (37.7%) and corresponds to the HLA-DR51 alleles; 122 peptides (4.4%) did not match to either group and were collected by the trash cluster.

In order to validate the solutions generated by GibbsCluster, we examined the composition of the clusters in terms of binding potential predicted by NetMHCIIpan-3.1 [47]. Both for the mono-allelic DR1 and poly-allelic DR15/51 serotypes discussed above, we obtained predicted percentile rank scores for all peptides in the cluster solutions and in their relative trash cluster (Figure 4). The predicted median rank score for HLA-DRB1*01:01 in the DR1 cluster was 4% (first quartile (Q1)=0.9, third quartile (Q3)=12), whereas the trash cluster had a median rank score of 41% (Q1=20.5, Q3=75). In the poly-allelic data, cluster 1 was associated with HLA-DRB1*15:01, and showed a median rank score of 13% (Q1=5, Q3=30); cluster 2 was associated to HLA-DRB5*01:01 and obtained a median rank score of 4% (Q1=1.1, Q3=11); peptides in the trash cluster were evaluated against both alleles, assigning the best rank of the two, which resulted in an average rank score of 41% (Q1=23, Q3=75) (Figure 4). Overall, the NetMHCIIpan percentile score distributions suggest that the trash cluster could successfully collect peptides with very poor correspondence to the known preferences of the MHC class II molecules, and that probably derived either from incorrect spectral matches or from contaminants. The relatively high predicted rank values for the peptides mapped to the HLA-DRB1*15:01 cluster further suggest that the binding motif for this molecule predicted by NetMHCIIpan-3.1, which was trained on binding affinity data, shared a rather weak overlap with the binding motif contained within the MS ligand data. This observation underlines the high potential of MS ligand data to complement our knowledge on peptide characteristics required for MHC antigen presentation, as previously remarked for MHC class I [2–4,21,23].

Generating prediction models from MS ligand data

The approaches described so far in this report are mainly concerned with extracting and visualizing meaningful patterns within complex, often noisy, mixtures of peptides sequences. A further step is the generalization of the motifs identified in the data at hand, by constructing prediction models. Machine learning algorithms such as NNAlign [32], when provided with

training examples suitably labeled (e.g. ligands vs. non-ligands), can be instructed to automatically learn the features that distinguish positive from negative examples. Such models can then be applied on external data sets to discover more occurrences of the patterns learned on the training data. In the context of peptide-MHC interactions, a good prediction model should have the ability to capture the binding preferences contained in the training data, both in terms of sequence motifs and peptide length distribution. In the next two sections, we illustrate some simple examples of prediction models directly constructed from MHC class I and class II eluted ligands.

MHC Class I prediction model

As an example application, we continue with the Abelin ligand elution dataset previously analyzed and filtered using GibbsCluster-2.0 (Figure 1). For each of the representative alleles HLA-A*68:02, HLA-B*35:01 and HLA-B*57:01, we prepared a training set consisting of post-filtering ligands (positive instances) and random natural peptides (negative instances). Positive instances were labeled with a target value of 1, negatives with a target value of 0. In line with earlier work [4], the amount of random negatives was imposed to be the same for each length 8 to 13, and corresponded for each length to five times the amount of positives for the most abundant peptide length. This uniform length distribution of the random negatives was adopted as a background against with machine learning can be employed to learn the amino acid and length preference of the natural binders.

On each of the three data sets, we trained a prediction model with the NNAlign-2.0 webserver, using the recommended preset options for MHC class I ligands of variable length. In a cross-validation experiment, the three models returned an area under the ROC curve (AUC) of 0.961, 0.984 and 0.979, respectively. In order to derive the amino acid and peptide length preferences learned by the model, we used it to evaluate a large set of 900,000 random natural peptides with a flat length distribution, and extracted the top 0.1% scoring peptides. The composition of these high-scoring peptides should reflect the main preferences identified by the method to distinguish positive from negative instances. Indeed, the binding motif drawn from the top 0.1% peptides closely reflects the amino acid preferences of the training data (Figure 5A-B). Moreover, all three methods could capture the preference for 9mer peptides over other peptide lengths; 10mers were moderately allowed, 8mers and 11mers were observed more infrequently (Figure 5C).

MHC Class II prediction model

To illustrate how the NNAlign framework can be used to construct MHC class II prediction models, we go back to the DR1 and DR15/51 data sets from Ooi et al. [44] previously filtered and clustered with GibbsCluster (Figure 3). To enrich the positive instances with artificial negative examples, a set of natural random negatives of length 11 to 19 amino acids was added to each eluted ligands data set. Positive instances were labeled with a target value of 1, negatives with a target value of 0. Similarly to the training set preparation described above for MHC class I, the amount of random negatives for each length corresponded to five times the amount of positives for the most abundant peptide length. For each of the three specificities deconvoluted by GibbsCluster in the DR1 and DR15/51 cells, we applied NNAlign-2.0 to generate a prediction model, using the preset parameters for MHC class II recommended by the NNAlign server. For the mono-allelic DR1 serotype, all ligands except those removed by the trash cluster were used to train a model. For the DR15/51 cells, for which the clustering analysis revealed two separate specificities, we generated a separate model from the ligands contained in each of the two clusters.

The three models revealed high internal consistency, with cross-validated performance of AUC=0.952, 0.974 and 0.952, respectively. NNAlign automatically generates a matrix (and logo) representation of the motif learned by the method, constructed from the top 1% scoring predictions from a large set of random natural peptides. We may compare the motifs learned by NNAlign to: *i)* the binding preferences in the MS training data, identified by GibbsCluster; *ii)* the GibbsCluster motifs identified in tetramer-validated epitopes extracted from the IEDB for the three DR molecules; *iii)* the binding preferences predicted by NetMHCIIpan-3.1 for these DR molecules. In general, the motifs learned by the NNAlign models share a remarkable overall correspondence to the preferences found by GibbsCluster for the MS ligand data, with similar amino acid enrichments at the anchor positions P1, P4 and P6, as well as the strong P9 for the DR51-associated ligands (Figure 6, first and second columns). Likewise, the binding motifs constructed from the rather small amount of tetramer-validated epitopes obtained from the Immune Epitope Database (IEDB) [48] for the three DR molecules (231 for HLA-DRB1*01:01, 129 for HLA-DRB1*15:01, 73 for HLA-DRB5*01:01) correspond well with the motifs of the NNAlign models, and the MS ligand data (Figure 6, third column). In contrast, the logos derived from *in-vitro* binding affinity data (NetMHCIIpan) in all cases show substantial differences to both the MS- and epitope-derived motifs (Figure 6, fourth column). These discrepancies are most pronounced for HLA-DRB1*15:01, where the NetMHCIIpan motif has weakly defined preferences at the anchor residues, and an

enrichment of arginine (R) throughout the binding motif; a preference that is completely absent from the MS and epitope-derived motif. Another, more subtle difference is the enrichment of glutamic acid (E) at P4 in the MS and epitope motifs for HLA-DRB1*01:01; this preference is absent in the NetMHCIIpan motif. Finally, NetMHCIIpan displays a preference for R/K at position P8 for HLA-DRB5*01:01; this anchor is completely absent in the motif derived from MS and tetramer-validated epitope data. Taken together, these results show that ligand elution is a stronger correlate of epitope presentation than peptide-MHC binding affinity, suggesting that epitope prediction models may greatly benefit from incorporating MS eluted ligand data.

Final remarks

The binding specificities of MHC molecules have been traditionally characterized using *in-vitro* assays of binding affinity. The peptide-MHC binding data amassed through decades of painstakingly low-throughput experiments have had a tremendous contribution to the characterization of the binding preference for the most prevalent MHC molecules, and more generally to the understanding of the peptide repertoire available for T cell recognition. However, because of the extreme polymorphism of the MHC-encoding genes, with up to several thousand allelic variants per locus, the full characterization of their specificities remains infeasible. Tandem mass-spectrometry has emerged in the past decade has a powerful, high-throughput alternative for the identification of peptides eluted on the surface of antigen-presenting cells.

The appeal of MS-based techniques does not only reside in the sheer amount of ligand data that can be detected in a single experiment. Because MS ligands are derived from a biological system that incorporates all properties of antigen presentation including binding affinity, binding stability, proper peptide processing and translocation, and impact of MHC binding chaperones, these techniques should capture additional signals besides the binding affinity measurable by *in-vitro* assays. Accurate tools for the identification of sequence motifs in eluted ligand datasets are essential to interpret the patterns underlying the immunopeptidome and to benefit from this data deluge.

In this report, we described some straightforward, efficient approaches to extract motifs from immunopeptidomes in a number of scenarios commonly encountered in the field. We outlined analyses for MHC class I and class II, both in cell lines expressing a single MHC allele and in unmodified cells with multiple MHC allelic variants. GibbsCluster [20] is our tool

of choice because it can effectively remove residual contaminants after FDR filtering, deconvolute multiple motifs in a mixture of peptides of variable length and because it works both for MHC class I and class II ligands. In general, MHC class I molecules have strong, well-defined motifs, and even in samples containing several specificities it is often feasible to separate them into individual clusters. An unresolved problem remains the unambiguous association of each cluster to individual MHC molecules, especially for alleles with unknown binding motifs. So far only Bassani et al. [27] have attempted to tackle this question, exploiting the co-occurrence of MHC class I alleles across different data sets of known haplotype to assign motifs to individual alleles. More work along these lines is needed to automatically annotate the MHC restriction of peptides in poly-allelic datasets. The current implementation of GibbsCluster assumes that each peptide is restricted to one and only one MHC molecule. When cells express different alleles with similar binding motifs, or in the case of MHC class II ligands binding to multiple alleles in different alignment frames, it is likely that an individual peptide can act as ligand for multiple MHCs in a mixture. Future improvements to the algorithm should aim to address this limitation and account for potential multiple restrictions of individual ligands.

Ultimately, prediction methods can only be as good as the data used to train them. While MHC ligands sequences obtained by mass-spectrometry show remarkable reproducibility and produce binding motifs consistent with those derived with more low-throughput assays, there remain several potential sources of error and bias in MS-based pipelines for ligand sequencing. For example, there is a documented underrepresentation of cysteine in MHC ligand data sets, as this amino acid interferes with MS precursor fragmentation [3,27]. Different software tools for spectrum-peptide mapping use different functions to score candidate sequences, and they will generally identify non-identical sets of ligands. Post-translational modifications (PTMs) have also been shown to have a role in shaping the MHC ligand repertoire [49].However, accounting for such modified residues further complicates accurate spectrum-peptide matching and PTMs are often not comprehensively considered in MS pipelines. Finally, common contaminants such as keratin and histone proteins are often co-eluted with MHC ligands and add a level of noise to the sequenced immunopeptidome [34,35]. Reducing biases and sources of error in the data-generation pipelines will also inevitably affect in a positive way the data interpretation and the prediction tools constructed on these data.

A number of recent reports have described the first prediction methods trained directly on MHC class I ligand elution data from MS [3,4,28,50]. Their results indicate that methods

trained on naturally presented peptides largely outperform prediction methods trained solely on *in-vitro* binding affinity data when it comes to identification of MHC ligands and epitopes. No reports have yet been published on models directly trained on MHC class II eluted ligands. Because the performance of MHC class II prediction methods still lags far behind their class I counterparts for epitope prediction, antigen processing factors are likely to play a major role in the generation of MHC class II ligands. Incorporating naturally processed ligand from MS experiments in the training pipelines of MHC class II prediction methods is an exciting and yet unexplored opportunity to close that gap. A simple but powerful approach to generate prediction models from ligand data is the NNAlign method [32]. We illustrate the construction of models from MS eluted ligands both for MHC class I and MHC class II, and show that they capture the preferences of the training data both in terms of binding motif and ligand length distribution. Taken together, these computational tools allow researchers to interpret motifs contained in immunopeptidomes and generate prediction models to scan protein databases for epitope candidates.

Acknowledgements

This work was supported by Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272201200010C; and by the Agencia Nacional de Promoción Científica y Tecnológica, Argentina (PICT-2016-0089).

The authors have declared no conflict of interest.

References

[1] Caron, E., Kowalewski, D.J., Chiek Koh, C., Sturm, T., et al., Analysis of Major Histocompatibility Complex (MHC) Immunopeptidomes Using Mass Spectrometry. *Mol. Cell. Proteomics MCP* 2015, 14, 3105–3117.

[2] Bassani-Sternberg, M., Gfeller, D., Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide-HLA Interactions. *J. Immunol. Baltim. Md 1950* 2016, 197, 2492–2499.

[3] Abelin, J.G., Keskin, D.B., Sarkizova, S., Hartigan, C.R., et al., Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* 2017, 46, 315–326.

[4] Jurtz, V.I., Paul, S., Andreatta, M., Marcatili, P., et al., NetMHCpan 4.0: Improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *bioRxiv* 2017.

[5] Ternette, N., Yang, H., Partridge, T., Llano, A., et al., Defining the HLA class I-associated viral antigen repertoire from HIV-1-infected human cells. *Eur. J. Immunol.* 2016, 46, 60–69.

[6] Yaciuk, J.C., Skaley, M., Bardet, W., Schafer, F., et al., Direct interrogation of viral peptides presented by the class I HLA of HIV-infected T cells. *J. Virol.* 2014, 88, 12992–13004.

[7] Berlin, C., Kowalewski, D.J., Schuster, H., Mirza, N., et al., Mapping the HLA ligandome landscape of acute myeloid leukemia: a targeted approach toward peptide-based immunotherapy. *Leukemia* 2015, 29, 647–659.

[8] Bassani-Sternberg, M., Coukos, G., Mass spectrometry-based antigen discovery for cancer immunotherapy. *Curr. Opin. Immunol.* 2016, 41, 9–17.

[9] Kalaora, S., Barnea, E., Merhavi-Shoham, E., Qutob, N., et al., Use of HLA peptidomics and whole exome sequencing to identify human immunogenic neo-antigens. *Oncotarget* 2016, 7, 5110–5117.

[10] Bassani-Sternberg, M., Bräunlein, E., Klar, R., Engleitner, T., et al., Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* 2016, 7, 13404.

[11] Cox, J., Mann, M., MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* 2008, 26, 1367–1372.

[12] Ma, B., Zhang, K., Hendrie, C., Liang, C., et al., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom. RCM* 2003, 17, 2337–2342.

[13] Zhang, J., Xin, L., Shan, B., Chen, W., et al., PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Mol. Cell. Proteomics* 2012, 11, M111.010587–M111.010587.

[14] Brosch, M., Yu, L., Hubbard, T., Choudhary, J., Accurate and sensitive peptide identification with Mascot Percolator. *J. Proteome Res.* 2009, 8, 3176–3181.

[15] Murphy, J.P., Konda, P., Kowalewski, D.J., Schuster, H., et al., MHC-I Ligand Discovery Using Targeted Database Searches of Mass Spectrometry Data: Implications for T-Cell Immunotherapies. *J. Proteome Res.* 2017, 16, 1806–1816.

[16] Delong, T., Wiles, T.A., Baker, R.L., Bradley, B., et al., Pathogenic CD4 T cells in type 1 diabetes recognize epitopes formed by peptide fusion. *Science* 2016, 351, 711–714.

[17] Liepe, J., Marino, F., Sidney, J., Jeko, A., et al., A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science* 2016, 354, 354–358.

[18] Rammensee, H.G., Friede, T., Stevanoviíc, S., MHC ligands and peptide motifs: first listing. *Immunogenetics* 1995, 41, 178–228.

[19] Andreatta, M., Lund, O., Nielsen, M., Simultaneous alignment and clustering of peptide data using a Gibbs sampling approach. *Bioinformatics* 2013, 29, 8–14.

[20] Andreatta, M., Alvarez, B., Nielsen, M., GibbsCluster: unsupervised clustering and alignment of peptide sequences. *Nucleic Acids Res.* 2017.

[21] Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L.J., Mann, M., Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics MCP* 2015, 14, 658–73.

[22] Probst, P., Kopp, J., Oxenius, A., Colombo, M.P., et al., Sarcoma Eradication by Doxorubicin and Targeted TNF Relies upon CD8(+) T-cell Recognition of a Retroviral Antigen. *Cancer Res.* 2017, 77, 3644–3654.

[23] Ritz, D., Gloger, A., Weide, B., Garbe, C., et al., High-sensitivity HLA class I peptidome analysis enables a precise definition of peptide motifs and the identification of peptides from cell lines and patients' sera. *Proteomics* 2016, 16, 1570–1580.

[24] Sofron, A., Ritz, D., Neri, D., Fugmann, T., High-resolution analysis of the murine MHC class II immunopeptidome. *Eur. J. Immunol.* 2016, 46, 319–328.

[25] Fugmann, T., Sofron, A., Ritz, D., Bootz, F., Neri, D., The MHC Class II Immunopeptidome of Lymph Nodes in Health and in Chemically Induced Colitis. *J. Immunol. Baltim. Md 1950* 2017, 198, 1357–1364.

[26] Mommen, G.P.M., Marino, F., Meiring, H.D., Poelen, M.C.M., et al., Sampling From the Proteome to the Human Leukocyte Antigen-DR (HLA-DR) Ligandome Proceeds Via High Specificity. *Mol. Cell. Proteomics MCP* 2016, 15, 1412–1423.

[27] Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., et al., Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput. Biol.* 2017, 13, e1005725.

[28] Nielsen, M., Connelley, T., Ternette, N., Improved prediction of Bovine Leucocyte Antigens (BoLA) presented ligands by use of MS eluted ligands and in-vitro binding data; impact for the identification T cell epitopes. *bioRxiv* 2017.

[29] Tenzer, S., Peters, B., Bulik, S., Schoor, O., et al., Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell. Mol. Life Sci. CMLS* 2005, 62, 1025–1037.

[30] Harndahl, M., Rasmussen, M., Roder, G., Dalgaard Pedersen, I., et al., Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *Eur. J. Immunol.* 2012, 42, 1405–1416.

[31] Andreatta, M., Schafer-Nielsen, C., Lund, O., Buus, S.S.S., Nielsen, M., NNAlign: A web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. *PLoS ONE* 2011, 6, e26781.

[32] Nielsen, M., Andreatta, M., NNAlign: a platform to construct and evaluate artificial neural network models of receptor-ligand interactions. *Nucleic Acids Res.* 2017.

[33] Sercarz, E.E., Maverakis, E., Mhc-guided processing: binding of large antigen fragments. *Nat. Rev. Immunol.* 2003, 3, 621–629.

[34] Hodge, K., Have, S.T., Hutton, L., Lamond, A.I., Cleaning up the masses: exclusion lists to reduce contamination with HPLC-MS/MS. *J. Proteomics* 2013, 88, 92–103.

[35] Mellacheruvu, D., Wright, Z., Couzens, A.L., Lambert, J.-P., et al., The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nat. Methods* 2013, 10, 730–736.

[36] Nielsen, M., Andreatta, M., NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med.* 2016, 8, 33.

[37] Rasmussen, M., Harndahl, M., Stryhn, A., Boucherma, R., et al., Uncovering the peptide-binding specificities of HLA-C: a general strategy to determine the specificity of any MHC class I molecule. *J. Immunol. Baltim. Md 1950* 2014, 193, 4790–4802.

[38] Nielsen, M., Connelley, T., Ternette, N., Improved Prediction of Bovine Leucocyte Antigens (BoLA) Presented Ligands by Use of Mass-Spectrometry-Determined Ligand and in Vitro Binding Data. *J. Proteome Res.* 2017.

[39] Nielsen, M., Lund, O., Buus, S., Lundegaard, C., MHC class II epitope predictive algorithms. *Immunology* 2010, 130, 319–328.

[40] Andreatta, M., Jurtz, V.I., Kaever, T., Sette, A., et al., Machine learning reveals a non-canonical mode of peptide binding to MHC class II molecules. *Immunology* 2017, 152, 255–264.

[41] Nielsen, M., Lundegaard, C., Worning, P., Hvid, C.S., et al., Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinforma. Oxf. Engl.* 2004, 20, 1388–1397.

[42] Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., et al., Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.* 1999, 17, 555–561.

[43] Andreatta, M., Nielsen, M., Characterizing the binding motifs of 11 common human HLA-DP and HLA-DQ molecules using NNAlign. *Immunology* 2012.

[44] Ooi, J.D., Petersen, J., Tan, Y.H., Huynh, M., et al., Dominant protection from HLA-linked autoimmunity by antigen-specific regulatory T cells. *Nature* 2017, 545, 243–247.

[45] Vogt, A.B., Kropshofer, H., Kalbacher, H., Kalbus, M., et al., Ligand motifs of HLA-DRB5*0101 and DRB1*1501 molecules delineated from self-peptides. *J. Immunol. Baltim. Md 1950* 1994, 153, 1665–1673.

[46] Scholz, E.M., Marcilla, M., Daura, X., Arribas-Layton, D., et al., Human Leukocyte Antigen (HLA)-DRB1*15:01 and HLA-DRB5*01:01 Present Complementary Peptide Repertoires. *Front. Immunol.* 2017, 8, 984.

[47] Andreatta, M., Karosiene, E., Rasmussen, M., Stryhn, A., et al., Accurate pan-specific prediction of peptide-MHC class II binding affinity with improved binding core identification. *Immunogenetics* 2015, 67, 641–650.

[48] Vita, R., Overton, J.A., Greenbaum, J.A., Ponomarenko, J., et al., The immune epitope database (IEDB) 3.0. *Nucleic Acids Res.* 2015, 43, D405–412.

[49] Van Els, C.A.C.M., Corbière, V., Smits, K., van Gaans-van den Brink, J.A.M., et al., Toward Understanding the Essence of Post-Translational Modifications for the Mycobacterium tuberculosis Immunoproteome. *Front. Immunol.* 2014, 5, 361.

[50] Giguère, S., Drouin, A., Lacoste, A., Marchand, M., et al., MHC-NP: predicting peptides naturally processed by the MHC. *J. Immunol. Methods* 2013, 400-401, 30–36.

Figure legends

Figure 1. Visualizing motifs and removing contaminants with GibbsCluster. A) Sequence motifs of three representative alleles before trash cluster filtering and **B)** after filtering. The post-filtering motifs have higher information content and lack the putative K/R contamination at P9. **C)** Distribution of NetMHCpan-3.0 percentile rank scores for peptides in the main cluster (red) and in the trash cluster (blue).

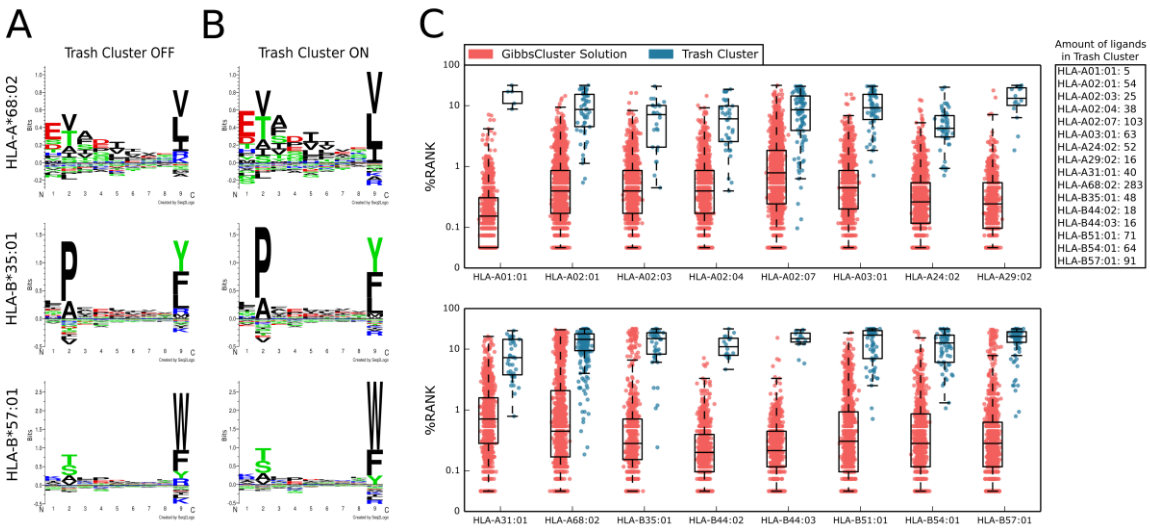


Figure 2. Clustering results for the HCC1143 cell line. The single cluster solution **(left)** is a mixture of multiple specificities, dominated by the most abundant alleles. The solution with highest information content corresponds to four clusters, with motifs highlighted in the red box **(center)**. The motifs identified by unsupervised clustering show a remarkable correspondence with those predicted by NetMHCpan-3.0 **(right)**. The GibbsCluster method was run using the default preset parameters for "MHC class I ligands of length 8-13", except for the number of iterations which was set to 100 (slower but more accurate), and number of groups, which was allowed to vary between 1 and 6. NetMHCpan logos were obtained from the NetMHCpan-3.0 website (http://www.cbs.dtu.dk/services/NetMHCpan-3.0/logos.php), and were constructed from the top 1% scoring peptides from a large set of natural random peptides.

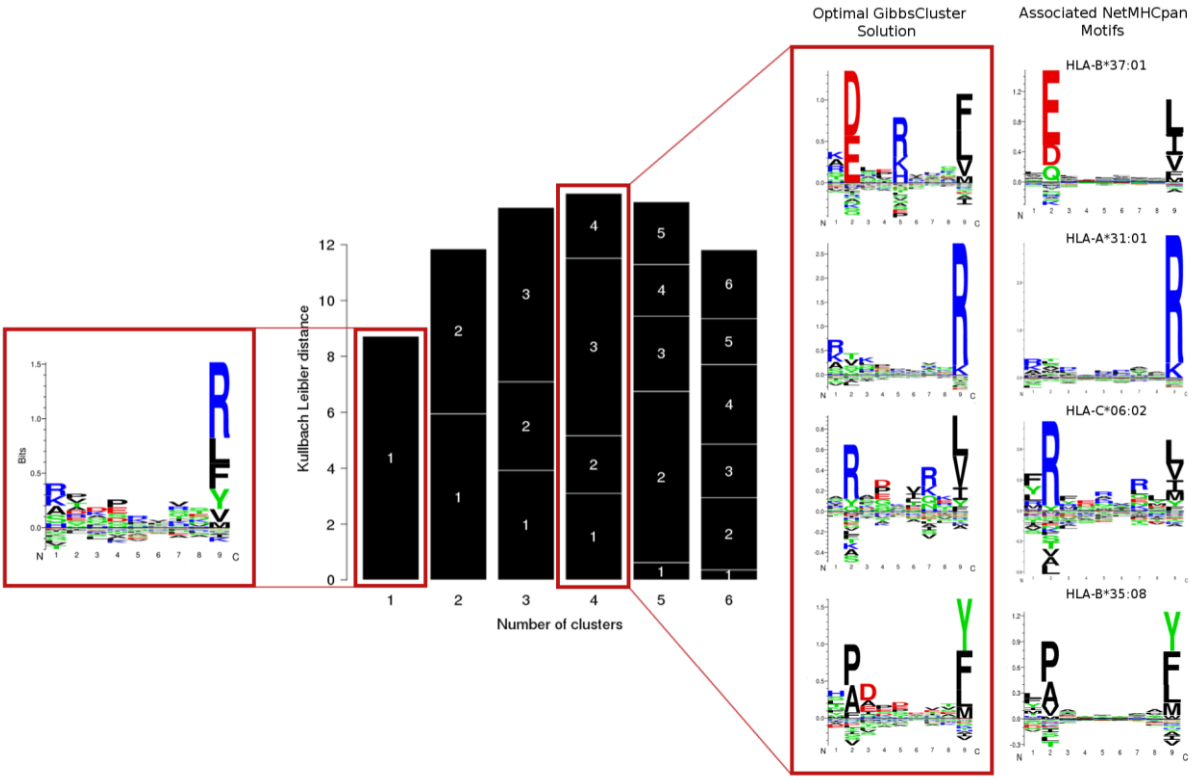


Figure 3. Sequence motifs identified by GibbsCluster-2.0 on MHC class II ligand data. The method identifies distinct amino acid preferences at the anchor positions P1, P4, P6 and P9 both without (left panels) and with (right panels) the trash cluster activated. **(A)** Visualizing the motif and removing outliers from the mono-allelic human-DR1 mouse-transfected cell lines. **(B)** Motif identification on mixed allelic data of DR15-DR51-EBV transformed cell lines.

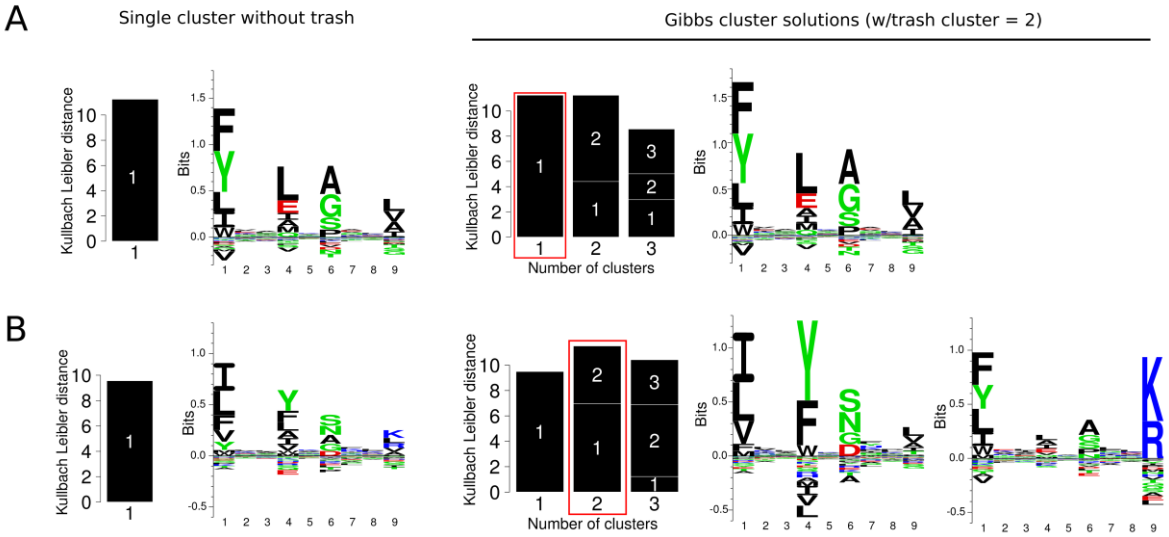


Figure 4. NetMHCIIpan percentile rank score for GibbsCluster solutions in the DR1 and DR15/51 data sets. Percentile rank scores were predicted by netMHCIIpan-3.1 for each GibbsCluster group with matching alleles present in MS data samples. In the case of the mixed allele dataset DR15/51, peptides in the trash cluster were scored by NetMHCIIpan to both DRB1*15:01 and DRB5*01:01, selecting the lowest rank score of the two.

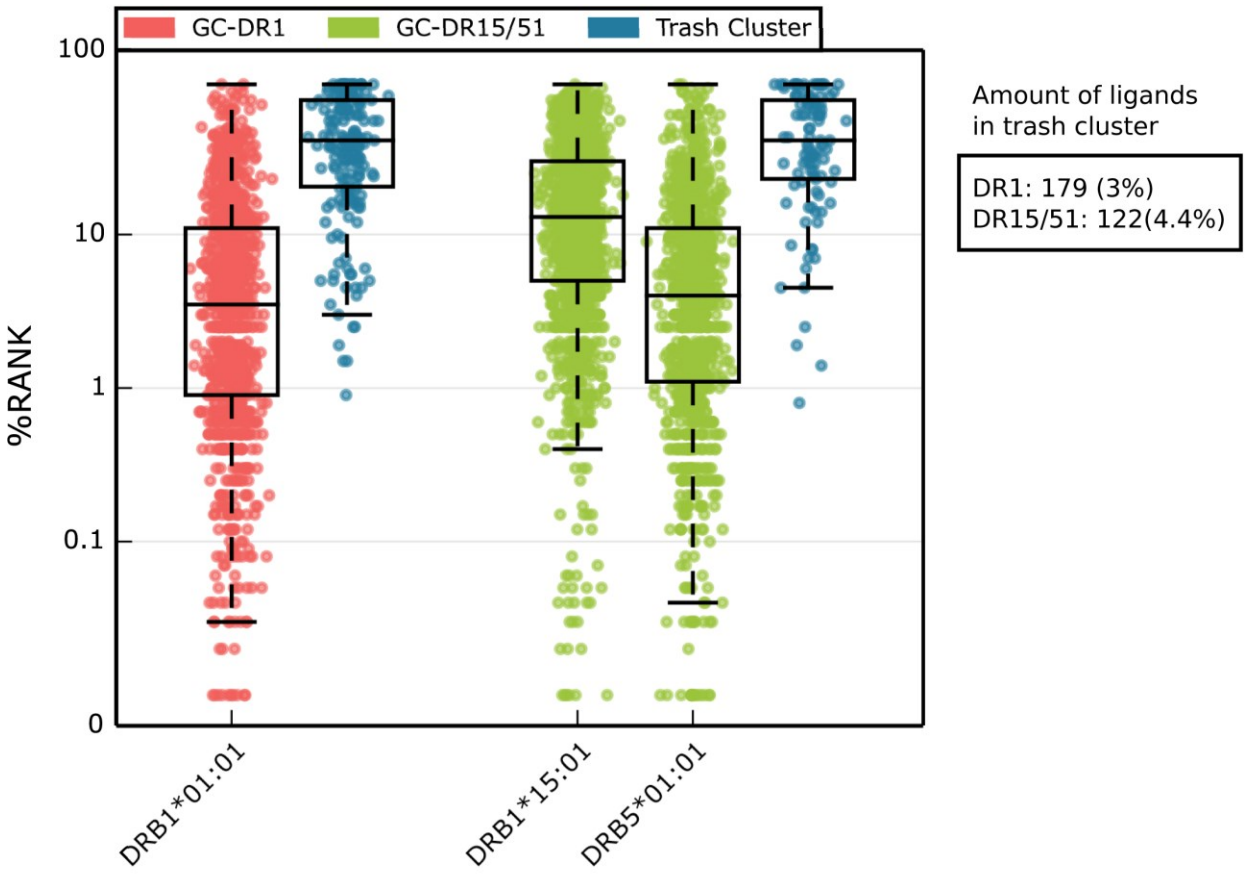


Figure 5. Generating prediction models from MS ligand data. A) Sequence motifs of the training data for three MHC class I alleles, aligned and filtered by GibbsCluster; **B)** Sequence motifs captured by NNAlign-2.0; **C)** Ligand length preferences in the training MS data compared to length preferences learned by the NNAlign model.

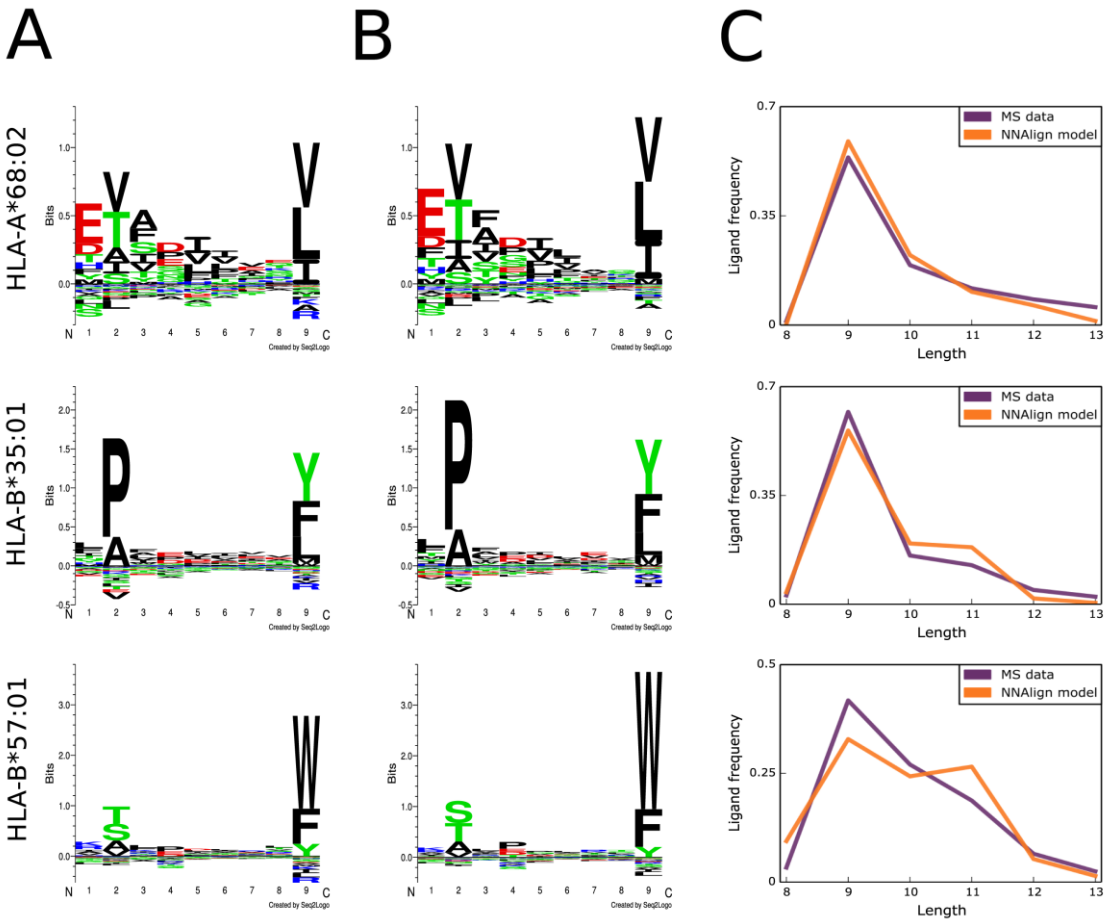
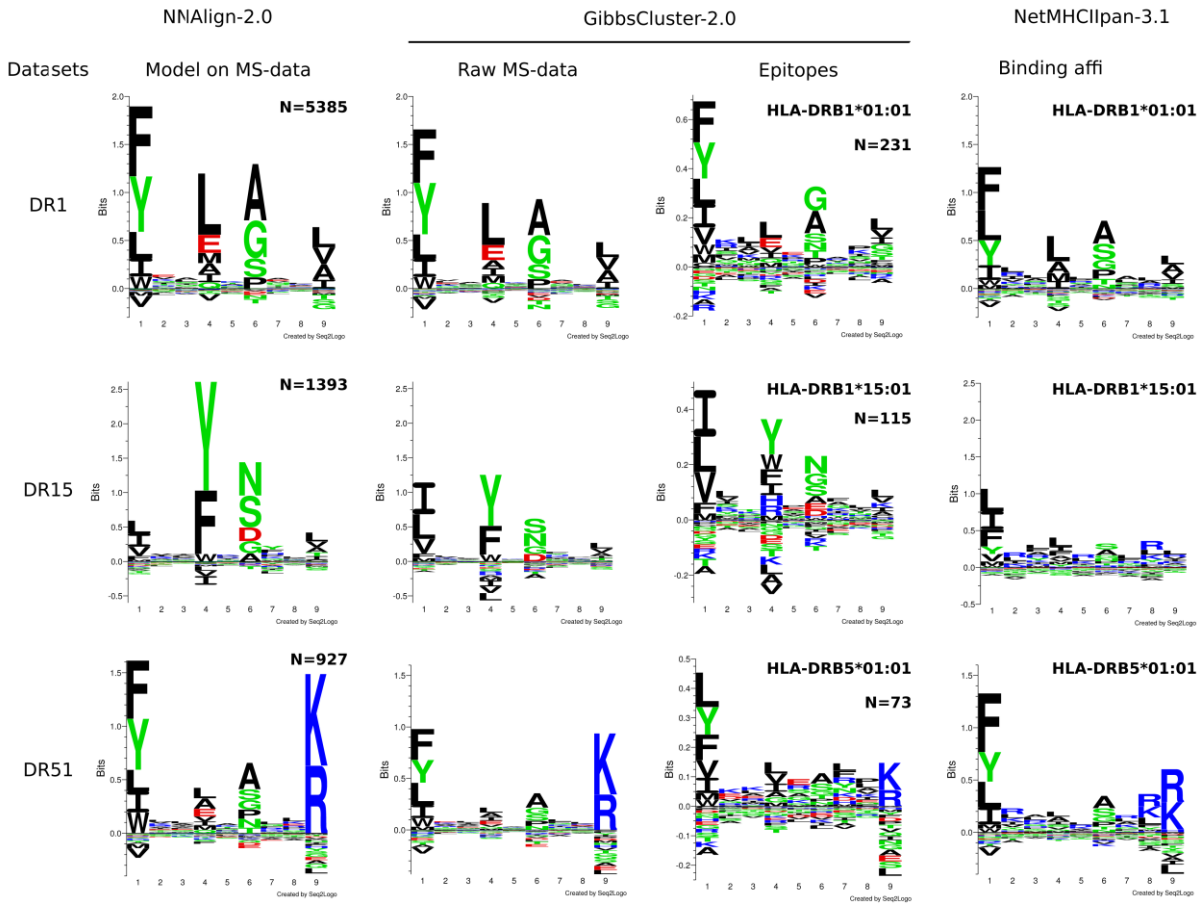


Figure 6. Comparison of motifs generated by different approaches for three HLA-DR alleles. NNAlign-2.0 motifs were obtained by training artificial neural networks on each MS data set, and evaluating 100,000 random peptides. The top scoring 1% peptides were used to build logos. Raw MS data were aligned, clustered and filtered in an unsupervised manner using GibbsCluster, with a trash cluster threshold = 2. The same procedure was applied to tetramer-positive data downloaded from the IEDB. Note that due to small data set size, epitope logos are shown in a different y-axis scale. Binding motifs for NetMHCIIpan-3.1 were determined by evaluating 100,000 random peptides, and visualizing the core motif of the top 1% scoring sequences.



Supplementary Figure S1. Sequence motifs of peptides collected by the main cluster and by the trash cluster for the 16 alleles in the Abelin data set.

Supplementary Figure S2. Sequence motifs of peptides collected by the trash cluster on the 7 alleles in the Bassani-Sternberg data set.

Supplementary Figure S3. Clustering of the HCC1143 cell lines by GibbsCluster (left) and NetMHCpan (right). Sequences were assigned by NetMHCpan to the allele in the haplotype with the lowest predicted %rank. If a peptide could be assigned to any MHC allele with rank<=2%, then it was discarded to the trash cluster. Note that, in this case, GibbsCluster could not deconvolute HLA-C*04:01 peptides.